

# Metric learning for phylogenetic invariants

Nicholas Eriksson

nke@stanford.edu

Department of Statistics,  
Stanford University,  
Stanford, CA 94305-4065

Yuan Yao

yuany@math.stanford.edu

Department of Mathematics,  
Stanford University,  
Stanford, CA 94305-4065

February 3, 2008

## Abstract

We introduce new methods for phylogenetic tree quartet construction by using machine learning to optimize the power of phylogenetic invariants. Phylogenetic invariants are polynomials in the joint probabilities which vanish under a model of evolution on a phylogenetic tree. We give algorithms for selecting a good set of invariants and for learning a metric on this set of invariants which optimally distinguishes the different models. Our learning algorithms involve linear and semidefinite programming on data simulated over a wide range of parameters. We provide extensive tests of the learned metrics on simulated data from phylogenetic trees with four leaves under the Jukes-Cantor and Kimura 3-parameter models of DNA evolution. Our method greatly improves on other uses of invariants and is competitive with or better than neighbor-joining. In particular, we obtain metrics trained on trees with short internal branches which perform much better than neighbor joining on this region of parameter space.

**Keywords:** Phylogenetic invariants, algebraic statistics, semidefinite programming, Felsenstein zone.

## 1 Introduction

Phylogenetic invariants have been used for tree construction since the linear invariants of Lake and Cavender-Felsenstein [Lak87, CF87] were found. Although the linear invariants of the Jukes-Cantor model are powerful enough to asymptotically distinguish between trees on 4 taxa [Lak87], these linear invariants do not perform well in simulations [Hue95].

In the last 20 years, the entire set of phylogenetic invariants has been found for many models of evolution (see [AR07] and the references therein). Since the invariants provide an essentially complete description of the model, using more invariants should give more power

to distinguish between different models. However, different invariants give vastly different power at distinguishing models and it is not known how to find the most powerful invariants.

In this paper, we use techniques from machine learning to find metrics on the space of invariants which optimize their tree reconstruction power for the Jukes-Cantor and Kimura 3-parameter phylogenetic models on trees with four leaves. Specifically, we apply *metric learning* algorithms inspired by [XNJR03] to find the metric which best distinguishes the models. For training data we use simulations over a wide range of the parameter space. Our main biological result is the construction of a metric which outperforms neighbor-joining on trees simulated from the Felsenstein zone (i.e., trees with a short interior edge). More generally, we find that metric learning significantly improves upon other uses of invariants and is competitive with neighbor joining even for short sequences and homogeneous rates.

Casanellas and Fernández-Sánchez [CFS06] also used the Kimura 3-parameter invariants to construct trees with four taxa. Their results indicated that invariants can sometimes perform better than commonly used methods (e.g., neighbor joining and maximum likelihood) for data that evolved with non-homogeneous rates and for extremely long sequences. They used the  $l_1$  norm on the space of invariants, weighing each polynomial equally.

This paper improves upon [CFS06] by showing how to improve upon the  $l_1$  norm on the space of invariants. The  $l_1$  norm behaves poorly since it weighs equally informative and non-informative invariants. Simulating data and using metric learning improves the performance of invariants by putting much more weight on the powerful invariants. This allows us to build an algorithm which is very accurate for trees with short internal edges.

We begin by briefly introducing the models and phylogenetic invariants we will use. Section 2 describes the metric learning algorithms; section 3 gives the results of our simulation studies; and we conclude with a short discussion.

By *phylogenetic tree*, we mean a binary, unrooted tree with labelled leaves. There are three such trees with four leaves labelled 0, 1, 2, 3, we call these trees  $T_1$ ,  $T_2$ , and  $T_3$  according to which leaf is on the same “cherry” as leaf 0. We consider two phylogenetic models on these trees: the Jukes-Cantor (JC69) model of evolution [JC69] and the Kimura 3-parameter (K81) model [Kim81], both with uniform root distribution.

These models associate to each edge  $e$  of the tree a transition matrix  $M_e = e^{Qt_e}$  where  $t_e$  is the length of the edge  $e$  and  $Q$  is a rate matrix:

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \cdot & \gamma & \alpha & \beta \\ \gamma & \cdot & \beta & \alpha \\ \alpha & \beta & \cdot & \gamma \\ \beta & \alpha & \gamma & \cdot \end{pmatrix}$$

for JC69 and K81 respectively, where  $\cdot = -\gamma - \alpha - \beta$ . For a given tree, we write  $p_{ijkl} = \Pr(\text{leaf } 0 = i, \text{leaf } 1 = j, \text{leaf } 2 = k, \text{leaf } 3 = l)$  for  $i, j, k, l \in \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$  and write  $p = (p_{\mathbf{AAAA}}, \dots, p_{\mathbf{TTTT}})$  for the joint probability distribution. *Phylogenetic invariants* are polynomial equations which are satisfied between the joint parameters. For example,  $p_{\mathbf{AAAA}} = p_{\mathbf{CCCC}}$  holds for both JC69 and K81, but since this equation is true for all three trees, will ignore it and similar equations.

**Example 1.** Consider the four-point condition on a tree metric [Bun74]. It says that if  $d$  is a tree metric on  $(ij : kl)$ , then

$$d_{ij} + d_{kl} < d_{ik} + d_{jl} = d_{il} + d_{jk}. \quad (1)$$

Given a probability distribution  $p$ , the maximum likelihood Jukes-Cantor distance is

$$d_{ij} = -\frac{3}{4} \log \left( 1 - \frac{4m_{ij}}{3} \right) \quad (2)$$

where  $m_{ij}$  is the fraction of mismatches between the two sequences, e.g.,

$$m_{12} = \sum_{w,x,y,z \in \{A,C,G,T\}, w \neq x} p_{wxyz}.$$

After substituting in (1) and exponentiating, the equality becomes

$$\left( 1 - \frac{4}{3}m_{ik} \right) \left( 1 - \frac{4}{3}m_{jl} \right) = \left( 1 - \frac{4}{3}m_{il} \right) \left( 1 - \frac{4}{3}m_{jk} \right). \quad (3)$$

This observation is originally due to Cavender and Felsenstein [CF87]. The difference of the two sides of (3) is a quadratic polynomial in the joint probabilities which we will call the four-point polynomial.

## 2 Methods

Since both models we consider are *group-based*, it is easiest to work in Fourier coordinates which can roughly be thought of as the  $m_{ij}$  coordinates in Example 1 (cf. [HP89, ES93, SS05]). The website <http://www.math.tamu.edu/~ljp/small-trees/> contains lists of invariants for different models on trees with a small number of taxa.

Our first task is building a set of invariants for the two models. The above website shows 33 polynomials (plus two implied linear relations) for JC69 and 8002 polynomials for K81. However, these sets of invariants are not closed under the symmetries of  $T_1$ . That is, each tree can be written in the plane in eight different ways (for example, the tree  $T_1$  can be written as  $(01 : 23)$ ,  $(10 : 23)$ ,  $\dots$ ,  $(32 : 10)$ ), and each of these induces a different order on the probability coordinates  $p_{ijkl}$ . We need a set of invariants which does not change under this reordering if we don't want the resulting algorithm to depend on the order of the input sequences.

After performing this calculation, we are left with 49 polynomials for JC69 and 11612 for K81. However, our metric learning algorithms run slowly as the number of invariants grows, so we had to find a subset of a more manageable size. We cut down the K81 invariants by testing each of the 11612 invariants individually on the entire parameter space and only keeping those which had good individual reconstruction rates. Specifically, we picked several different values for  $\gamma, \alpha, \beta$  and kept only those invariants which gave over a 62% reconstruction rate individually for sequences of length 100. The result of this calculation is sets of invariants  $\mathbf{f}_i^{JC69}$  and  $\mathbf{f}_i^{K81}$  of cardinality 49 and 52.

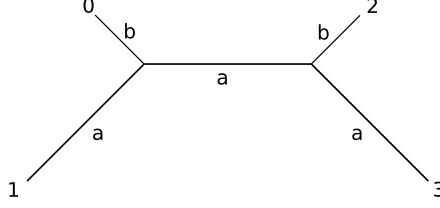


Figure 1: The tree used in the simulations. Branch lengths  $a$  and  $b$  ranged from 0.01 to 0.75 in intervals of 0.02.

**Definition 2.** Given a probability distribution  $\hat{p}$ , and invariants  $\mathbf{f}_i = (f_{i,1}, \dots, f_{i,n})$ , for tree  $T_i$  (for  $i = 1, 2, 3$ ), let

$$\mathbf{f}_i(\hat{p}) = (f_{i,1}(\hat{p}), \dots, f_{i,n}(\hat{p}))$$

be the point in  $\mathbb{R}^n$  obtained by evaluating the invariants for  $T_i$  at  $\hat{p}$ .

If the probability distribution  $\hat{p}$  actually comes from the model  $T_1$ , then we will have  $\mathbf{f}_1(\hat{p}) = 0$ ,  $\mathbf{f}_2(\hat{p}) \neq 0$ , and  $\mathbf{f}_3(\hat{p}) \neq 0$  generically (that is, except for points  $\hat{p}$  which lie on the intersection of two or more models). This fact suggests that we can just pick the tree  $T_i$  such that  $\mathbf{f}_i$  is closest to zero. However, the next example shows that it is quite important to pick good polynomials and weigh them properly.

**Example 3.** Figure 2 shows the distribution of four of the invariants from  $\mathbf{f}_i^{JC}$  on data from simulations of 1000 i.i.d. draws from the Jukes-Cantor model on  $T_1$  over a varying set of parameters. The histograms show the distributions for the simulated tree ( $T_1$ ) in yellow and the distributions for the other trees in gray and black.

Polynomial 10 (upper left) distinguishes nicely between the three trees with the correct tree tightly distributed around zero. It is correct 97% of the time on our space of trees (Figure 1). Polynomial 48 (upper right) also shows power to distinguish between all three trees, but the distributions are much more overlapping — it is only correct 50.8% of the time. Polynomial 10 is the four-point invariant from Example 1, polynomial 48 is one of Lake’s linear invariants. The two other examples show a polynomial (23) which is biased towards selecting the wrong tree (only 16% correct), and a polynomial (45) for which the correct tree is tightly clustered around zero, but the incorrect trees are indistinguishable and have wide variance (88.9% correct).

The parameters used for the simulations are described in Figure 1. Since 1000 samples should be quite enough to determine the structure of a tree on four taxa, it is revealing that many of the individual polynomials are quite poor (the mean prediction rate for all 49 polynomials is only 42%). The invariants have quite different variances and means and it is not optimal to take each one with equal weight.

This example shows that we need to scale and weigh the individual invariants. Recall that for a positive (semi)definite matrix  $A$  the Mahalanobis (semi)norm  $\|\cdot\|_A$  is defined by

$$\|x\|_A = \sqrt{x^t A x}.$$

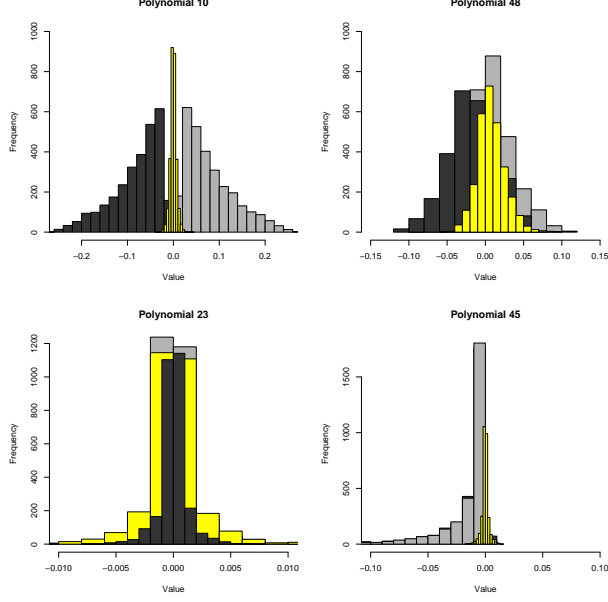


Figure 2: Distributions of four polynomials  $f_{i,10}$ ,  $f_{i,48}$ ,  $f_{i,23}$  and  $f_{i,45}$  on simulated data. The yellow histogram corresponds to the correct tree, the black and gray are the other two trees.

Notice that since  $A$  is positive semidefinite, it can be written as  $A = UDU^t$  where  $U$  is orthogonal and  $D$  is diagonal with non-negative entries. Thus the square root  $B = U\sqrt{D}U^t$  is unique. Now since  $\|x\|_A^2 = x^t A x = (Bx)^t (Bx) = \|Bx\|^2$ , we can view learning such a metric as finding a transformation of the space of invariants that replaces each point  $x$  with  $Bx$  under the Euclidean norm. Accordingly, we will be searching for a positive semidefinite matrix  $A$  on the space of invariants which is “optimal”.

Let  $\hat{p}(\theta)$  be an empirical probability distribution generated from a phylogenetic model on tree  $T_1$  with parameters  $\theta$ . We wish to find  $A$  such that the condition

$$\|\mathbf{f}_1(\hat{p}(\theta))\|_A < \min(\|\mathbf{f}_2(\hat{p}(\theta))\|_A, \|\mathbf{f}_3(\hat{p}(\theta))\|_A)$$

is typically true for most  $\hat{p}(\theta)$  chosen from a suitable parameter space  $\Theta$ .

Now suppose that  $\Theta$  is a finite set of parameters from which we generate training data  $\mathbf{f}_1(\hat{p}(\theta)), \mathbf{f}_2(\hat{p}(\theta)), \mathbf{f}_3(\hat{p}(\theta))$  for  $\theta \in \Theta$ . As we saw above, each of the eight possible ways of writing each tree induces a signed permutation of the coordinates of each  $\mathbf{f}_i(\hat{p}(\theta))$ . We write these permutations in matrix form as  $\pi_1, \dots, \pi_8$ . Given this training data, we wish to solve the following optimization problem.

**Algorithm 4. (Metric learning for invariants)**

*Input:* model invariants  $\mathbf{f}_i$  for  $T_i$  and a finite set  $\Theta$  of model parameters.

*Output:* a semidefinite matrix  $A$

*Procedure:*

1. Reduce the sets  $\mathbf{f}_i$  of invariants to a manageable size by testing individual invariants on data simulated from  $\theta \in \Theta$ .

2. Augment the resulting sets so that they are closed under the eight permutations of the input which fix tree  $T_1$ .
3. Compute the signed permutations  $\pi_1, \dots, \pi_8$  which are induced on the invariants  $\mathbf{f}_1$  by the above permutations.
4. Solve the following semidefinite programming problem:

$$\begin{aligned}
&\text{Minimize:} && \sum_{\theta \in \Theta} \xi(\theta) + \lambda \text{tr} A \\
&\text{Subject to:} && \|X_1(\hat{p}(\theta))\|_A^2 + \gamma \leq \min(\|X_2(\hat{p}(\theta))\|_A^2, \|X_3(\hat{p}(\theta))\|_A^2) + \xi(\theta), \\
&&& \pi_i A = A \pi_i \quad \text{for } 1 \leq i \leq 8, \\
&&& A \succeq 0, \quad \text{and} \\
&&& \xi(\theta) \geq 0,
\end{aligned}$$

where  $A \succeq 0$  denotes that  $A$  is a positive semidefinite matrix.

5. Alternatively, if we restrict  $A$  to be diagonal, this becomes a linear program and can be solved for much larger sets of invariants and parameters.

In the optimization step, we use a regularization parameter  $\lambda$  to keep  $A$  small and a margin parameter  $\gamma$  to increase the margin between the distributions. This is a convex optimization problem with a linear objective function and linear matrix equality and inequality constraints. Hence it is a semidefinite programming (SDP) problem. The SDP problem above has a unique optimizer and can be solved in polynomial time. Its complexity depends on the capacity of the set  $\Theta$  since each point in  $\Theta$  contributes a linear constraint.

For the range of parameters we consider in Section 3, we use  $\#(\Theta) = 1444$ . In our experiments to solve the SDP we use SeDuMi 1.1 [Stu99] or DSDP5 [BY05] with YALMIP [LÖ4] as the parser. Matlab code to implement the above algorithm can be found at <http://math.stanford.edu/~yuany/metricPhylo/matlab/>.

We found that although SeDuMi often runs into numerical issues, it generally finds a good matrix  $A$  with competitive performance to the neighbor-joining algorithm. DSDP is better in dealing with numerical stability at the cost of more computational time. We have found that setting  $\lambda = 0.0001$  and  $\gamma = 0.005$  gives good results in our situation. For example, in the case of the JC69 model with a  $49 \times 49$  semidefinite matrix  $A$ , YALMIP-SeDuMi takes 55.7 minutes to parse the constraints and solve the SDP, while YALMIP-DSDP takes 167.1 minutes to finish the same job. For details on experiments, see the next section.

Our algorithm was inspired by some early results on metric learning algorithms such as [XNJR03] and [SSSN04], which aim to find a (pseudo)-metric such that the mutual distances between similar examples are minimized while the distances across dissimilar examples or classes are kept large. Direct application of such an algorithm is not quite suitable in our setting. As shown in Figure 2, the correct tree points are overlapped by the two incorrect trees. For points in the overlapping region, it is hard to tell whether to shrink or stretch their mutual distance. However, when the points appear in triples, it is possible that for each triple the one closest to zero is generated from the correct tree. Our algorithm is based on such an intuition and proved successful in experiments.

After using Algorithm 4 to find a good metric  $A$ , the following simple algorithm allows us to construct trees on four taxa.

**Algorithm 5. (Tree construction with invariants)**

*Input:* A multiple alignment of 4 species and a semidefinite matrix  $A$  from Algorithm 4.

*Output:* A phylogenetic tree on the 4 species (without branch lengths).

*Procedure:*

1. Form empirical distributions  $\hat{p}$  by counting columns of the alignment.
2. Form the vectors  $\mathbf{f}_i = (f_{i,1}(\hat{p}), \dots, f_{i,n}(\hat{p}))$  for  $1 \leq i \leq 3$ .
3. Return  $T_i$  where the vector  $\mathbf{f}_i$  has smallest  $A$ -norm  $\|\mathbf{f}_i\|_A = \sqrt{\mathbf{f}_i^t A \mathbf{f}_i}$ .

### 3 Results

We tested our metric learning algorithms for the invariants  $\mathbf{f}^{JC69}$  and  $\mathbf{f}^{K81}$  as described above. We trained two metrics for JC69 on the tree in Figure 1. The first used simulations from branch lengths between 0.01 and 0.75 on a grid with increments of magnitude 0.02, for a total of 1444 different parameters. This region of parameter space was chosen for direct comparison with [Hue95, CFS06]. The second metric used parameters  $0.01 \leq a \leq 0.25$  and  $0.51 \leq b \leq 0.75$  with increments of 0.01, giving 625 trees with very short interior edges. Similarly for K81, we learned a metric using parameters  $0.01 \leq a, b \leq 0.75$  and several sets of  $\gamma, \alpha, \beta$ . After learning metrics, we performed simulation tests on the same space of parameters that was used to train. We compared neighbor-joining [SN87], phylogenetic invariants with the  $l_1$  or  $l_2$  norm and phylogenetic invariants with our learned norms.

Since the edge lengths are large for part of the parameter space, we often see simulated alignments with more than 75% mismatches between pairs of taxa. In such a case, (2) returns infinite distance estimates under the Jukes-Cantor model. So the results depend on how neighbor joining treats infinite distances. In PHYLIP [Fel04], the program `dnadist` doesn't return a distance matrix if some distances are infinite. However, in PAUP\*, infinite distances are set to a fixed large number. Since we are only concerned with the tree topology, we believe that the most fair comparison is between phylogenetic invariants and the method of PAUP\*. However, it should be noted that this can make a major difference in results using neighbor joining, since often the correct tree can be returned even if some distances are infinite. See Figure 3 for an example of the difference and be warned that comparison between simulations studies done in different ways is difficult.

Table 1 shows the results of 100 simulations at each of the 1444 parameter values for various sequence lengths using the JC69 model. It gives the percent correct over all 144,400 trials for five different methods: invariants with  $l_1$ ,  $l_2$ , and  $A$ -norms and neighbor joining (using Jukes-Cantor distances and allowing infinite distances). The contour plots in Figure 3 show how the reconstruction rates vary across parameter space for the five methods for a sequence length of 100. Notice that the  $A$ -norm shows particularly good behavior over the

Full parameter space					Felsenstein zone				
Length	$l_1$	$l_2$	$A$ -norm	NJ	Length	$l_1$	$l_2$	$A$ -norm	NJ
25	62.7	59.8	74.5	75.5	25	31.8	31.3	58.9	52.5
50	71.9	66.3	85.0	85.9	50	35.9	33.0	69.9	63.2
75	76.7	69.6	90.0	90.4	75	39.2	34.4	76.5	69.5
100	79.8	72.0	92.7	92.9	100	42.2	35.8	81.2	73.9
200	86.4	77.6	97.0	96.6	200	50.9	39.1	90.1	83.5
300	89.2	80.1	98.2	97.7	300	55.4	40.4	93.6	87.8
400	91.1	82.1	98.7	98.2	400	59.5	41.4	95.1	90.2
500	92.3	83.5	99.0	98.4	500	62.4	42.5	95.6	91.4

Table 1: Percent of trials reconstructed correctly for the Jukes-Cantor model over the entire parameter space and the Felsenstein zone for the respective metrics.

$(\gamma, \alpha, \beta) = (0.1, 3.0, 0.5)$					$(\gamma, \alpha, \beta) = (0.2, 0.5, 0.3)$				
length	$l_2$	$l_2$ restrict	A	NJ	length	$l_2$	$l_2$ restrict	A	NJ
25	59.2	66.9	71.5	62.7	25	63.7	67.4	70.9	65.0
50	68.0	77.5	82.1	72.7	50	71.6	77.5	81.1	74.2
75	73.4	82.7	86.9	79.3	75	75.4	82.7	86.4	80.5
100	76.8	85.8	89.7	82.6	100	77.6	85.8	89.3	83.4
200	84.6	90.8	94.3	90.1	200	83.3	91.3	94.1	89.7
300	88.2	92.6	93.1	95.6	300	86.4	93.2	95.7	91.8
400	90.1	93.6	96.4	94.8	400	88.5	94.3	96.5	93.0
500	91.4	94.2	96.9	95.7	500	90.0	93.7	96.3	93.2

Table 2: Percent of trials reconstructed correctly for the Kimura 3-parameter model over the entire parameter space for two choices of  $\gamma, \alpha, \beta$ .

entire range of parameters, even in the “Felsenstein zone” in the upper left corner. When trained on the Felsenstein zone, the learned metric can perform even better. Table 1 shows the result of training a metric on this zone. Notice that the  $A$ -norm is now quite a bit better than neighbor joining, even though the  $l_1$  and  $l_2$  norms are terrible. However, this learned norm is slightly worse on the whole parameter space than the metric trained on the whole space.

Table 2 shows results for the K81 model under two choices of  $(\gamma, \alpha, \beta)$ . We only report the  $l_2$  scores, since the  $l_1$  scores are similar. Of note is the column “ $l_2$  restrict” which shows the  $l_2$  norm on the top 52 invariants as ranked by individual power on simulations as in the previous section. This column is better than the  $l_2$  norm on all 11612 invariants, showing that many invariants are actually harmful. The  $A$ -norm again improves on even the restricted  $l_2$  and beats neighbor-joining (run with K81 distances) on all examples.



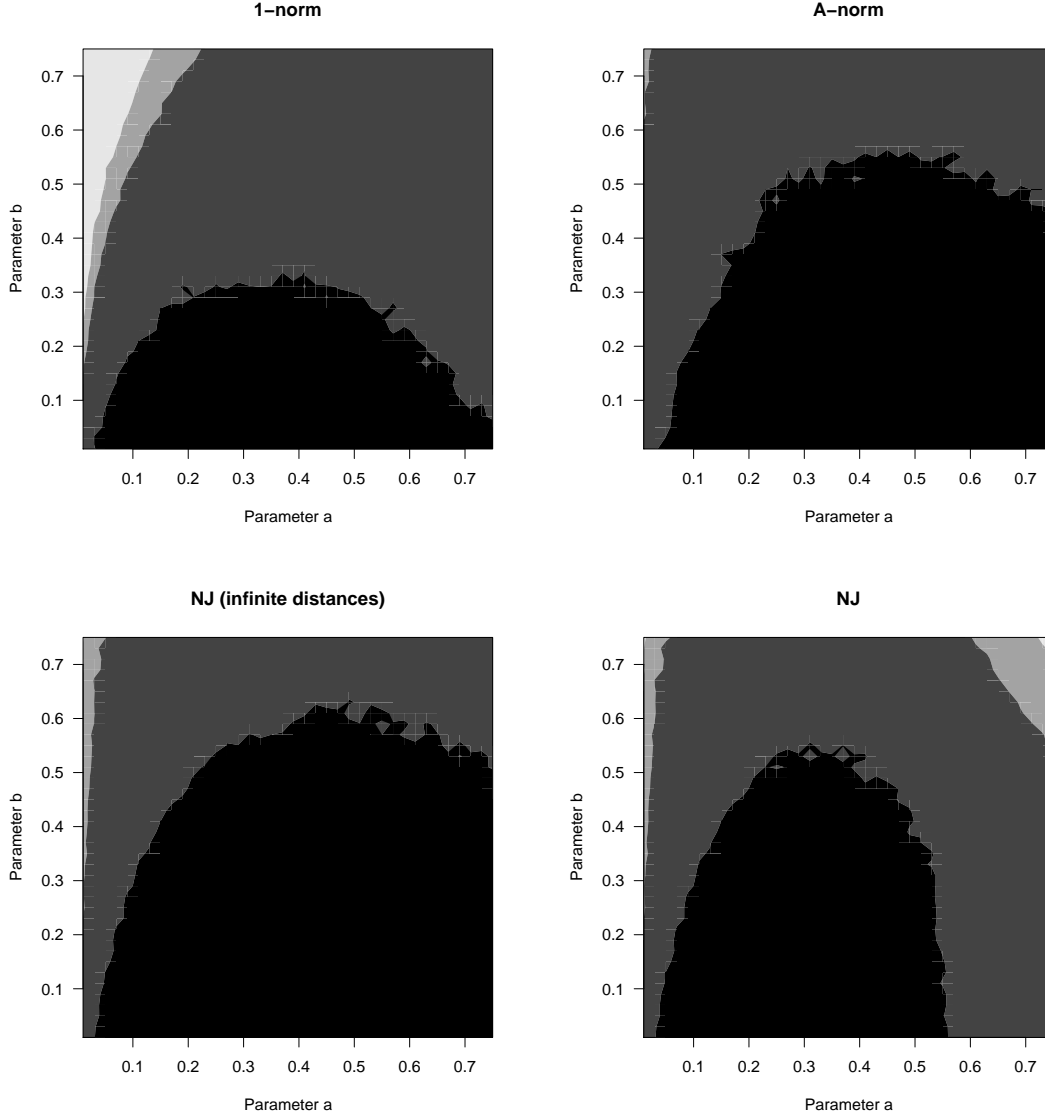


Figure 3: Contour plots for the three reconstruction methods for the Jukes-Cantor model over parameter space with alignments of length 100. Black areas correspond to parameters  $(a, b)$  for which the tree was reconstructed correctly over 95% of the time, gray for over 50%, light gray for over 33%, and white for under 33%.

## 4 Discussion

We have shown that machine learning algorithms can substantially improve the tree construction performance of phylogenetic invariants. As an example, for sequences of length 100, the four-point invariant (Example 1) for the K81 model is correct 82% of the time on data simulated from K81 with parameters (0.1, 3.0, 0.5). This is quite a bit better than the  $l_2$  norm on all 11612 invariants (76.8%, Table 2).

The paper [CFS07] describes an algebraic method for picking a subset of invariants for the K81 model. They reduce to 48 invariants which give an improvement over all 11612 invariants (up to 82.6% on the above example using the  $l_2$  norm). However, of these 48, only 4 of them are among the top 52 we selected for  $\mathbf{f}^{K81}$ , and the remaining 44 invariants are mostly quite poor (42% average accuracy). After taking the closure of these 48 invariants, there are 156 total and the performance actually drops to 78.3%. It seems that the conditions for an invariant to be powerful are not particularly related to the algebraic criterion used in [CFS07].

All invariant based methods heavily depend on the set of invariants that we begin with. Learning diagonal matrices  $A$  had mixed performance, which further suggests that the generating set we are using for the invariants is non-optimal. We believe that it is an important mathematical problem to understand what properties are shared by the good invariants. We suggest that symmetry might be an important criterion to construct other polynomials like the four-point condition with good power.

The learned metrics in this paper are somewhat dependant on the parameters chosen to train them. This can be a benefit, as it allows us to train tree construction algorithms for specific regions of parameter space (e.g., the Felsenstein zone). However, we hope that improvements to the metric programming will allow us to train on larger parameter sets and thus obtain uniformly better algorithms.

Notice that these methods only recover the tree topology, not the edge lengths. We believe that if the edge lengths are needed, they should be estimated after building the tree, in which case standard statistical methods such as maximum likelihood can be used easily. While the invariants discussed in this paper may not be practical for large trees, we believe there is great use in understanding fully the problem of building trees on four taxa. For example, these methods can either be used as an input to quartet-based tree construction algorithms or as a verification step for larger phylogenetic trees.

For a method of building trees on more than four taxa using phylogenetic invariants, see [Eri05, KKPP06], which use numerical linear algebra to evaluate invariants given by rank conditions on certain matrices in order to construct phylogenetic trees. This amounts to evaluating many polynomials at once, allowing it to run in polynomial time.

The matrices  $A$  for JC69 and K81 used in the tests can be found at <http://stanford.edu/~nke/data/metricPhylo>. A software package that can run these tests is available at the same website. It includes a program for simulating evolution using any Markov model on a tree and several programs using phylogenetic invariants.

## Acknowledgments

N. Eriksson was supported by NSF grant DMS-0603448 and wishes to thank MSRI and the IMA for their hospitality. Y. Yao was supported by DARPA grant 1092228. We wish to thank E. Allman, M. Drton, D. Ge, F. Memoli, L. Pachter, J. Rhodes, and Y. Ye for helpful comments and especially G. Carlsson for his encouragement and computational resources.

## References

- [AR07] E Allman and J Rhodes. Molecular phylogenetics from an algebraic viewpoint, 2007. To appear.
- [Bun74] Peter Buneman. A note on the metric properties of trees. *J. Combinatorial Theory Ser. B*, 17:48–50, 1974.
- [BY05] Steven J. Benson and Yinyu Ye. DSDP5: Software for semidefinite programming. Technical Report ANL/MCS-P1289-0905, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, September 2005. Submitted to ACM Transactions on Mathematical Software.
- [CF87] J Cavender and J Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71, 1987.
- [CFS06] M. Casanellas and J. Fernández-Sánchez. Performance of a New Invariants Method on Homogeneous and Non-homogeneous Quartet Trees. *Mol Biol Evol*, page msl153, 2006.
- [CFS07] M. Casanellas and J. Fernández-Sánchez. Geometry of the Kimura 3-parameter model, 2007. available at arXiv:math.AG/0702834.
- [Eri05] Nicholas Eriksson. Tree construction using singular value decomposition. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for Computational Biology*, chapter 19, pages 347–358. Cambridge University Press, Cambridge, UK, 2005.
- [ES93] S Evans and T Speed. Invariants of some probability models used in phylogenetic inference. *The Annals of Statistics*, 21:355–377, 1993.
- [Fel04] J Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, 2004.
- [HP89] M Hendy and D Penny. A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38(4), 1989.

- [Hue95] John P. Huelsenbeck. Performance of phylogenetic methods in simulations. *Sys Biol*, 1(44):17–48, 1995.
- [JC69] TH Jukes and C Cantor. Evolution of protein molecules. In HN Munro, editor, *Mammalian Protein Metabolism*, pages 21–32. New York Academic Press, 1969.
- [Kim81] M Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences, USA*, 78:454–458, 1981.
- [KKPP06] Young Rock Kim, Oh-In Kwon, Seong-Hun Paeng, and Chun-Jae Park. Phylogenetic tree constructing algorithms fit for grid computing with SVD. Available at <http://arxiv.org/abs/q-bio.QM/0611015>, 2006.
- [Lö4] J. Löfberg. YALMIP : a toolbox for modeling and optimization in MATLAB. In *Computer Aided Control Systems Design*, pages 284–289, 2004.
- [Lak87] JA Lake. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4:167–191, 1987.
- [SN87] N Saitou and M Nei. The neighbor joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.
- [SS05] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *J Comput Biol*, 12(4):457–481, May 2005.
- [SSSN04] Shai Shalev-Shwartz, Yoram Singer, and Andrew Y. Ng. Online learning of pseudo-metrics. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.
- [Stu99] J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999. Special issue on Interior Point Methods (CD supplement with software).
- [XNJR03] Eric Xing, Andrew Y. Ng, Michael Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2003.